HOW TO CRITIQUE AN ARTICLE ON THERAPY

DR. BRUCE F. WALKER D.C., M.P.H.*

Abstract: The ability to critique the literature and source out relevant information to answer a clinical question is a skill that is being introduced into the under-graduate curricula of most health professions. Posing a clinical question on therapy, sourcing the literature, reviewing critically what you find and then hopefully answering the question is central to the evidence-based method. The very foundation of clinical teaching and clinical practice will in the future rely on the "evidence-based method". A checklist is the easiest and quickest way to review journal articles.

Key Indexing Terms: Therapy, chiropractic, evidence-based method, clinical epidemiology, evidence-based chiropractic, checklist.

INTRODUCTION

The ability to critique the literature and source out relevant information to answer a clinical question is a skill that is being introduced into the under-graduate curricula of most health professions. Posing a clinical question, sourcing the literature, reviewing critically what you find and then hopefully answering the clinical question is generically known as "Evidence-based health care". In our profession I prefer to see the label "Evidence-based chiropractic".

Generally speaking journal articles which address clinical questions fall into one of four categories:

1. Therapy
2. Diagnostic tests
3. Causation
4. Prognosis

In this article I deal exclusively with therapy. Therapy means any treatment applied to gain a healing or preventive response. However, it should be noted that an article on prevention has slightly different review criteria than the ones found below.

A checklist is the easiest and quickest way to review journal articles. A checklist (Figure 1) has been constructed from a number of key texts (1- 5). Read the checklist and if you have difficulty with any section review the explanatory notes below.

* Private Practice
Suite 16, Hyde Park Centre,
Hyde Park, Queensland. AUSTRALIA 4812.

HOW TO UNDERSTAND AND USE THE CHECKLIST

1. a) Advantages of randomisation

There is no bias in the allocation to treatments groups. This evens out the groups and removes potential biases.

Study groups will tend to be comparable with respect to all variables except for the interventions being studied. Baseline characteristics, and confounding factors will be evenly distributed.

Randomisation is extremely important. If there was no randomisation discard the article. It is worth noting there are several methods of randomisation, look for this detail in the paper.

b) Research design

Is the research design the correct method of studying the question asked? It may be that the question could have been answered with a more appropriately designed study.

This can usually be answered after a critical appraisal of the study and a basic working knowledge of research design. For example, the strengths and weaknesses of randomised controlled trials, cohort studies, case-control studies, cross-sectional studies, case series and so on.

2. a) Patient selection

The study group should be a representative sample of the wider population of those with the condition. Therefore, it is important that those selected are not a biased group in some manner. (Known as "Selection bias"). For example, in the case of chronic low back pain it would be wrong to have too many males, smokers, labourers, aged persons etc. The population sampled should be found in the methods and results section.

b) Patient follow up

All patients should be accounted for at the end of the trial. Drop outs need to be identified and an explanation of their fate given. Patients do not usually drop out of a study for trivial reasons. They may refuse to have the therapy, have got worse, have been cured or simply moved town. All of these reasons are important factors which may effect the results. If the drop outs are not discussed adequately, then establish how many there are, add them to the treatment failure list and re-calculate the results. Similarly add them to the

Figure 1.

CHECKLIST AND WORKSHEET AS A GUIDE FOR EVALUATING AN ARTICLE ABOUT THERAPY

Article title: _____

_____

Authors: _____

Publication details: _____

		Circle		
1.	a) Was the assignment of patients to treatment randomised?	yes	no	?
	b) Was the research design fundamentally sound ?	yes	no	?
	IF SO PROCEED TO Q2, IF NOT DISCARD STUDY!			
2.	Were all patients entered into the trial properly accounted for and attributed at its conclusion? ie.			
	a) Was patient selection appropriate?	yes	no	?
	b) Was follow up complete?	yes	no	?
	c) Were patients analysed in the groups to which they were assigned?	yes	no	?
3.	Were the patients blinded to which treatment they received? sb	yes	no	?
	Were clinicians blinded to treatments given? db	yes	no	?
	Were other key study personnel blinded as well? obs. bias	yes	no	?
4.	Were both groups similar at baseline? ie			
	a) Demographic data (table?)	yes	no	?
	b) Prognostic factors (stratified?)	yes	no	?
	c) General health questions?	yes	no	?
	If not, were the statistics adjusted to account for this?	yes	no	?
5.	a) Was the eligibility criteria for the entry to the trial appropriate?	yes	no	?
	b) Was the sample size calculated before trial commencement?	yes	no	?
6.	Were all spectrums of disease represented in the sample?	yes	no	?
7.	Were the interventions used appropriate ie.			
	a) Were they sensible and described adequately?	yes	no	?
	b) Are they affordable?	yes	no	?
	c) Are they available?	yes	no	?
	d) Was there a placebo used?	yes	no	?
	e) Was there a non-treatment group (natural history)?	yes	no	?

Page 1 of 2

8.	Were the groups "treated equally".	yes	no	?
9.	Was there any co-intervention?	yes	no	?
	Was there any contamination?	yes	no	?
	Did patients comply with treatment regimes and instructions?	yes	no	?
10.	Were the instruments of measurement used:			
	a) reasonable and adequate for all clinically important outcomes?	yes	no	?
	b) reliable and valid	yes	no	?
11.	Was normal defined?	yes	no	?
12.	a) Was there sufficient follow up?	yes	no	?
	b) Were adverse effects documented?	yes	no	?
13.	How large was the treatment effect? ie			
	a) Was it statistically significant?	yes	no	?
	b) Was it clinically significant?	yes	no	?
	c) Was the difference biologically plausible?	yes	no	?
14.	How precise was the estimate of effect? ie.			
	a) Was sufficient sample size used?	yes	no	?
	b) Were confidence intervals given?	yes	no	?
	c) Was there any data dredging?	yes	no	?
15.	Were the limitations of the study discussed adequately?	yes	no	?
16.	Were there any other biases operating, and if so in what direction?	yes	no	?
17.	Can the results be applied to my patient care?	yes	no	?
18.	a) Is the treatment efficacious?	yes	no	?
	b) Is it effective?	yes	no	?

Conclusion: _____

_____

Page 2 of 2

"cured" list and do likewise. If this process does not alter the outcome of the trial then you may overlook the drop outs.

c) Intention to treat analysis

Patients should be analysed as being part of the group to which they were originally allocated during the randomisation process. This is known as "intention to treat analysis". This strategy preserves the integrity of the randomisation process. For example if we exclude non-compliant patients (even those who did not receive any treatment) we leave behind a group who may have been destined to do better than drop outs. Intention to treat analysis prevents systematic exclusion of drop outs from the results, thereby better reflecting the real life outcome of the intervention.

3. Blinding

a) A single blind trial is where the patient is unaware of the treatment they are receiving, e.g. placebo or drug. A double blind trial has patient blinding and in addition the treating doctor is also blinded to the treatment received. This is relatively simple with drugs but impossible with manual therapy such as manipulation.

However, it is important to blind the assessors of outcome and initial assessment in all trials of this nature. If not, these assessors may inadvertently influence the outcome by comments made to the patient, or a preference for a particular intervention. This is called observer bias.

The lack of blinding in trials of non-drug therapies imposes an extra burden on the researcher to ensure that those responsible for the delivery of the treatment are competent and able to provide a consistent level of care. For example, in a trial comparing manipulation with mobilisation, it would be important for the practitioners to be equally skilled at their craft and have equal enthusiasm for the intended outcome.

4. Baseline data

a) Similarity of all treatment groups is important. The authors should provide a table of demographic data demonstrating that all treatment groups had

similar demographics. Of course, randomisation usually takes care of this in large trials. In the case of smaller trials some special block randomisation techniques can be used. Usual baseline data should include, age, gender and socio-economic status.

b) Prognostic factors are other variables which should be equal between the groups. For instance, in a trial on back pain, both groups should at baseline have a similar mean disability score and chronicity. Otherwise one groups prognosis may be worse at the outset and bias the outcome.

c) General health status of groups should be similar. This can be established with validated health questionnaires such as the Sickness Impact Profile. There should not be a significant difference between groups in general health scores.

If there are differences found between groups, certain statistical analysis can be performed to adjust for this, for example multi-variate analysis. It is sufficient to ascertain that this has been done, unless you are of course a statitiscal wiz.

5. a) Eligibility criteria. Of those entered into the trial, were those included an appropriate sample capable of answering the research question posed by the authors? Were there too many exclusions? On some occasions there are so many exclusions that the sample bears little resemblance to the population likely to reasonably present for therapy in your own setting. Note that it is worth checking the authors conclusion to ascertain whether it is suitably worded to take into consideration the sample finally selected. For example, in a trial of manipulation and back pain conducted on adults, the conclusion should specifically mention adults as the study group. The study cannot be generalised to children.

b) Sample size calculations should be performed before the study commences and mention of how the figure was derived should be found in the methods section of the paper. The probability of detecting a difference between groups when one actually exists is called the power of the test. Increasing the sample size increases the power.

6. All spectrums of disease should ideally be included in the study, i.e. mild, moderate and severe. If, for example, only severe disease is included, the results may make the therapy results appear artificially poor. However, if the purpose of the study is to look only at a particular stage of disease (e.g. chronic) then clearly this should be mentioned and the conclusion confined to this stage.

7. Were the interventions used appropriate?

a) Were they sensible and described adequately? The authors should in their introduction make out a case for the use of the therapies under study. The use of the therapy should be inherently sensible, biologically plausible and ethically sound.

All therapies should be defined clearly so that another researcher could repeat the study and have the therapies performed based on the original description. A fastidious trial uses only "pure" and consistent forms of treatment, e.g. aspirin vs. manipulation vs. acupunture, while a pragmatic trial uses broader definitions e.g. chiropractic vs. medicine vs. physiotherapy. Regardless of the approach a detailed definition of either should be available in the methods section.

b) All therapies should be affordable, or if not then of sufficient importance that it would be likely that the cost could be reduced with mass introduction (for example, hepatitis B vaccination), or the cost met by Government (for example, simvastatin for hypercholesterol-aemia).

c) Is the therapy available? The therapy under consideration should be generally available or likely to be so.

d) The use of a placebo in trials of therapy is considered important, but is controversial. The advantage is that placebo gives a base measure of improvement by just providing help to the subject. It may be (for instance) that the therapies under study all deliver the same level of improvement to the subjects and that this level is the same as the placebo group. How can the reader tell unless a placebo group is included?

On the other hand there is the ethical question of withholding therapy from those in the trial, and by doing so possibly prolonging their pain and suffering.

The epidemiologist in me says include the placebo, while the few humanitarian bones that I have left say leave it out. What do you think?

e) A non-treatment group receives no therapy (placebo or active). The inclusion of such a group is worthwhile to demonstrate the natural history of the disease. The same arguments arise as in the placebo question.

However, such groups, whether placebo or non-treatment, are important in research to establish baseline characteristics of the disease and to decide whether therapy is useful.

8. Were the groups treated equally? It is important to ascertain whether the study groups were treated equally. For example, did the clinicians and staff display the same amount of optimism for all therapies. Did they all receive a reasonable amount of treatment specific to the therapy? For example, enough manipulation or NSAID over a reasonable period of time to be fair to the therapy. Even simple matters such as the availability of appointments for the various therapies should be equal.

9. The three C's.
Co-intervention, contamination, compliance.

a) Co-intervention. After randomisation, patients may receive a variety of interventions other than the ones being studied. If these occur unequally in the two groups and affect outcomes, they can introduce bias. If very effective non-study treatments are allowed at the physicians discretion, co-intervention may be a big problem. All such co-interventions allowed should be mentioned in the "Methods" section and should be minimal in effect. For instance, if it was found that in a trial on low back pain comparing manipulation and mobilisation that one group were taking more NSAID's than the other, one might be very worried about the effect.

b) Contamination. This occurs when the control group accidentally receives the active treatment.

c) Compliance with therapy is very important, and the researchers should give detail of how they established compliance, particularly where home compliance is required for example, drugs or exercises.

10. Instruments of measurement

Were they reasonable and adequate for all clinical outcomes? Measuring health and aspects of health is a difficult task. Many instruments of measurement have been devised but few have been found to be reliable and valid. In therapy research, look to ascertain that the instruments used have been found to be reliable (inter-rater and intra-rater) and valid. Also establish in your own mind that they are the appropriate measures to answer the question posed in the research. It should be noted that reliability does not mean validity. Reliability is the extent to which repeated measurements of a relatively stable phenomenon fall closely to each other, while validity is the degree to which the results of a measurement correspond to the true state of the phenomenon being measured. For instance, palpation of the spine might be highly reliable, but it may not correspond at all to the existence of a manipulable lesion (validity).

11. Was normal defined?

In a trial measuring blood pressure, it is imperative that normal ranges are known, so that abnormal can also be defined. Normal should be known from the outset and defined in the study methods section. For instance, in a trial involving the measurement of straight leg raising, is 60 or 90 degrees normal? Or in a trial measuring leg length discrepancy, should we allow 5 mm or 5 cm tolerance as being within normal limits? As you can see normal is not always an easy value to define.

12. Follow up

a) Was there sufficient follow up?
Terminating a trial before significant changes take place constitutes a serious flaw in research design. For example, low back pain research has often been criticised because intervention, and particularly measurement of outcomes, was terminated to early. In the latter case, the question often raised being "Is there a long lasting effect from the treatment or is it only short lived?"

b) Adverse effects
The detailing of adverse effects in trials on therapy is essential for the reader to be able to judge benefit against risk. All adverse effects should be reported along with their frequency, severity and duration. However, it must be recognised that therapy trials are not good estimates of the risk of treatment. They usually underestimate risk, particularly the less common side effects. Large cohort studies are the design of choice.

13. How large was the treatment effect?

a) Was it statistically significant?
There is often a need to quantify the degree to which chance variability may account for the observed treatment results in a research study. There are many statistical tests which are appropriate for the particular situation. Common tests are the *t*-test and chi-square. A commonly reported measure of statistical significance is the probability value or p-value. By convention if the p-value is less than 0.05 then the test is significant. This means that there is less than or equal to a 1 in 20 chance of the observed value being observed

by chance alone. Remember though that 1 in 20 times this test will be wrong and show that an observed value that is statistically significant, when in fact it is not. This is why repeated research on the same question is often required. More on this later in question 14c). In brief, statistical significance is important in showing that the values observed in the study on therapy were (on the balance of probability) either different from group to group or not. To illustrate, in a hypothetical trial comparing manipulation and mobilisation for the treatment of acute back pain, it was found that the manipulation group were back at work in a mean time of 3.2 days whereas the mobilisation group were back at work in 3.6 days. Can we say that such a difference is statistically significant? This will in fact depend on the number of subjects in the trial. If only 10 subjects were in each group it is unlikely, however if there were 100 in each group it is very likely. Statistical significance is important, always look for it. Of course it is not always necessary. If the difference in recovery time in the example above was 3.2 days and 32 days, one hardly needs a statistical tests to demonstrate a difference between reasonably sized groups.

b) Was it clinically significant?

An outcome may be statistically significant but not clinical significant. For instance, in the example in 13a) above, if the mean recovery time was 3.2 days and 3.24 days respectively and the trial involved a large number of subjects, say 3000, then it is likely that the resultant difference of 0.04 days would reach statistical significance, but it is doubtful that such a small difference could be said to be clinically significant.

A meaningful difference in treatment effect should be defined by the authors at the outset.

c) Biological plausibility?

Even if there is a statistically and clinically significant difference found in the results of the study, the reader is well advised to consider whether the results are actually biologically plausible. There may be some other factors at play that are not immediately obvious giving a spurious result.

14. How precise was the estimate of effect?

a) Was the sample size sufficient?

The estimate of sample size is based on a few facts and a relatively straight forward mathematical calculation. Determination of sample size is also known as power analysis or determining the power of the study. Suffice to say that the greater the sample size the greater the power of the study.

b) Were confidence intervals given?

Confidence intervals are statistical values which estimate the variability of the result. They define an upper limit and a lower limit with an associated probability. The most commonly used confidence interval is that associated with a 95% probability.

Let us illustrate this rather abstract notion with an example. In a fictitious study, serum calcium was measured before and after manipulation of the spine. Forty three male patients were tested and the mean pre-manipulative serum calcium was measured at 9.9 mg/dL, with a standard deviation of 0.66. The mean of the post-manipulation group was 9.5 mg/dL. The researchers wanted to know whether the results in this group of manipulated subjects were in fact different from the normal population (pre-manipulated state). The 95% confidence limits were calculated and showed that we can be 95% confident that the true post manipulation mean falls between 9.3 and 9.7 mg/dL. So the conclusion is that the post manipulative group probably fall outside the population pre-manipulation mean and are indeed different.

c) Was there any data dredging?

Data dredging is the performance of a myriad of hypothesis tests on a data set after the study has been completed in the hope of finding some result that is statistically significant. The safest way to perform such post hoc tests is to reduce the p-value commensurate with the number of post-hoc tests performed. Remember from 13a) above that, based on the 0.05 p-value, 1 in every 20 tests will provide a spuriously false positive result. Therefore, by dividing the 0.05 value by the number of extra post hoc tests we can be assured that the chances of a spurious result are not increased.

For example, if 100 post hoc tests are to be performed, the p-value will change from 0.05 to 0.0005. Always check for data dredging. Authors should have established what tests were to be performed prior to commencement of the trial.

15. Were the limitations of the study discussed?

It is important that the authors discuss adequately the limitations of their study. Such limitations should then accord with the wording of the conclusion. A major limitation in the study (even if admitted) may render the conclusion invalid.

16. Were there any other biases operating, and if so in what direction?

There are many potential biases in therapy research. If you find one then determine whether the bias operates to make the results appear better or worse, and judge for yourself what potential impact this may have on the conclusion.

17. Can the results be applied to my patient care?

Therapy research conducted in a tertiary setting such as a hospital may attract a very different type of patient to those who attend a local clinic. Careful analysis of the patient demographics, prognostic data and general health indices will assist in determining equivalence. The point is, the patients who were in the trial may bear little resemblance to your own, therefore generalising the results to your own clinic may be invalid.

18. Is the treatment efficacious? Is it effective?

A therapy may be shown in a research study to be efficacious. That is under the circumstances it was trialled it has been shown to work. However, it may not be effective in the real world. For instance, epidural injections may be shown to be effective for the treatment of chronic low back pain, but if the general public by and large do not want to be injected into the spine then the treatment is not effective. Further, if chiropractic treatment was ever to shown to be efficacious in the treatment of asthma, would the public at large choose chiropractic treatment over the use of cheaper, convenient and more readily available drugs.

CONCLUSION

Now, go to it and try reviewing the next therapy paper (preferably a randomised controlled trial) that interests you. Having performed your critical appraisal of the paper, try and conclude whether on balance you are inclined to accept the conclusions drawn by the authors. Does it actually answer the research question posed?

REFERENCES

1. Sackett DL, et al. Clinical epidemiology. 2nd ed. Batton: Little, Brown and Co., 1991.
2. Fletcher RH, Fletcher SW, Wagner EH. Clinical epidemiology: the essentials. 2nd ed. Baltimore: Williams and Wilkins, 1988.
3. Greenberg RS. Medical epidemiology. Lange. Connecticut. 1993.
4. Hennekens CH, Buring JE. Epidemiology in Medicine. Batton: Little Brown and Co., 1987.
5. Friedman GD. Primer of epidemiology. 3rd ed. McGraw Hill. New York. 1987
6. Dawson-Saunders B, Trapp RG. Basic clinical and biostatistics. Lange, Connecticut, 1990.